In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institution shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

_____

A STUDY OF FLOATING POINT ARITHMETIC

A THESIS

Presented to

The Faculty of the Graduate Division

By

Thomas Woodward Bookhart

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

September, 1964

A STUDY OF FLOATING POINT ARITHMETIC

Approved:

[illegible]

[illegible]

[illegible]

Date approved by Chairman: 11/23/64

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. William J. Kammerer, my thesis advisor, for his guidance and help in the preparation of this thesis. I also wish to express my gratitude to Dr. George C. Caldwell and Dr. David L. Finn for reading the manuscript and offering helpful suggestions in improving it. Special thanks are also due Mrs. Peggy Weldon for her excellent typing of the thesis.

TABLE OF CONTENTS

INTRODUCTION

Since applied mathematics and engineering ultimately are reduced to numerical results, their user must be able to employ the numbers and formulas encountered so as to obtain the best results. It is necessary in making numerical calculations to replace numbers which have an infinite representation with rational numbers of finite length, whether in computation by hand or with the aid of an electronic computer. Therefore, in making numerical calculations, only a finite subset of the real numbers is actually used. In this study certain subsets of the real numbers which are used in digital computers will be examined.

The early digital computers were of the fixed point type in which the radix point was fixed for the computations. This presented the time-consuming task in programming of arranging the calculations so that all quantities concerned remained within the limits of the machine and yet were expressed to an accuracy sufficient to insure the desired precision in the results. Difficulty in programming was eliminated by the introduction of a computer which allowed numbers to be expressed in a floating radix form. Numbers of this form are represented by $d \cdot \beta^e$. The first machine of this kind was the Bell Telephone Laboratories' relay computer model V. This was a decimal machine $(\beta = 10)$ in which $0.1 \leq |d| < 1$, $-19 < e < 19$, and d was expressed to an accuracy of seven figures. Today most of the electronic digital computers are of the floating point type.

In Chapter I certain collections of numbers used in floating

point computers will be studied. In addition the pseudo operations of addition and multiplication of floating point numbers will be defined. Since these operations are determined by the design of the particular machine, the operations as defined in this study may differ slightly from those of some computers. A set of floating point numbers, along with the pseudo operations, will be called a floating point number system. Properties of the floating point number system will be examined and compared with those of the real number system. Also in Chapter I a normalized floating point number system will be examined. In this system all floating point numbers except zero have a nonzero leading digit in the fractional part. After a pseudo operation which produces a zero in the first digit of the fractional part, the number is normalized or shifted until the first digit is different from zero. Properties of the normalized floating point number system will also be compared with those of the real number system.

In Chapter II the pseudo operations of addition, multiplication, and division are compared with the corresponding real number operations and bounds on the error created by performing pseudo operations in lieu of real number operations are established. Error analysis in digital computers was pioneered for fixed point machines by Goldstein and Von Neumann [5] and Householder [6, 7]. A similar analysis for floating point machines was done by Carr [3] and most of the results of Chapter II appear in the paper by Carr.

It is shown in Chapter I that the floating point number systems lack many of the good properties of the real number system. Included in these properties is the absence of associative and distributive laws.

Chapter III deals with this problem. The order in which numbers are pseudo added and multiplied is examined to determine which arrangement gives rise to the smallest error bound.

The study closes (Chapter IV) with a brief study of a special problem, that of finding the roots or a second degree polynomial using polynomial deflation. The order in which the roots are found is discussed (Wilkinson [14] discusses this problem for a general polynomial).

CHAPTER I

NUMBER SYSTEMS

A. The Real Number System

The purpose of this study is not to make a detailed analysis of the real number system but rather to study certain subsets of the real numbers which are used to approximate it for computational purposes. Certain properties of the real number system will be compared with those of the other systems. To facilitate this study some of the properties of the real number system will be listed in this section. For a more detailed discussion of the real numbers see, for instance, Landau [8].

Properties of the Real Number System

1. Between any two distinct real numbers, there is another real number.

2. For any real number a, there is another real number b such that $a < b$.

3. A measure of distance between two real numbers a and b is given by

$$d(a, b) = |a - b| .$$

For any set of real numbers a, b and c this measure of distance satisfies

(i) $d(a, b) \geq 0$

(ii) $d(a, b) = 0$ if and only if $a = b$

(iii) $d(a, b) = d(b, a)$

(iv) $d(a, b) \leq d(a, c) + d(c, b)$.

The two basic operations, addition and multiplication, which are defined on the real numbers, satisfy

4. Addition.

(i) Every two real numbers, x and y, have a unique sum $x + y$.

(ii) The commutative law holds. That is, if x and y are any real numbers, $x + y = y + x$.

(iii) The associative law holds. That is, if x, y, and z are any real numbers, $(x + y) + z = x + (y + z)$.

(iv) There exists a unique real number 0 such that for any real number x, $x + 0 = x$.

(v) Corresponding to any real number x there exists a unique real number $-x$ such that $x + (-x) = 0$.

5. Multiplication.

(i) Every two real numbers, x and y, have a unique product $x \cdot y$.

(ii) The commutative law holds. That is, if x and y are any real numbers, $x \cdot y = y \cdot x$.

(iii) The associative law holds. That is, if x, y, and z are any real numbers, $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.

(iv) There exists a unique real number $1 \neq 0$ such that for any real number x, $x \cdot 1 = x$.

(v) Corresponding to every real number $x \neq 0$ there exists a unique real number x^{-1} such that $x \cdot x^{-1} = 1$.

6. Distributive Law

If x, y, and z are any real numbers, $x \cdot (y+z) = x \cdot y + x \cdot y$.

Let R denote the set of all real numbers and let $R^* = \{\text{all } x \in R \mid x \neq 0\}$. Property 4 implies that $\{R, +\}$ is an Abelian group. Property 5 states that $\{R^*, \cdot\}$ is an Abelian group. Properties 4, 5, and 6 together imply that the real number system, $\{R, +, \cdot\}$, is a field.

B. Floating Point Number System

In making numerical calculations either by hand or with the aid of an electronic computer, practicality dictates that only a finite subset of the real numbers be used. In this section certain subsets which are used for computational purposes in digital computers will be defined and discussed.

In most electronic computers used for scientific computation, numbers are represented in the computer by a sequence of digits and an exponent. The allowable number of digits and the limits on the exponent are determined by the particular computer word length. These numbers are called floating point numbers.

Definition 1.1: Floating Point Number. Let M, L, U, and β be given integers such that $L < U$ and $\beta > 1$. A floating point number is a real number of the form $\pm \left(\sum_{i=1}^{M} d_i \beta^{-i} \right) \beta^e$, where d_i and e are integers such that $0 \leq d_i < \beta$ and $L < e < U$.

Thus a floating point number has the form $d \cdot \beta^e$, where d is an M-digit number, $|d| < 1$, and e is restricted to certain bounds. The word length of the particular computer would determine the number of

digits, M, and the exponent limits, L and U.

As an illustration, let $\bar{F}$ denote the set of all floating point numbers for $M = 4$, $\beta = 10$, $L = -5$, and $U = 5$. Then any member of $\bar{F}$ is of the form

$$\pm \cdot d_1 d_2 d_3 d_4 \; 10^e = \pm \left(\frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \frac{d_4}{10^4} \right) 10^e$$

where d_i $(i = 1,2,3,4)$ and e are integers which satisfy

$$0 \leq d_i < 10 \qquad i = 1,2,3,4$$
$$-5 < e < 5 \; .$$

This floating point representation possesses a number of troublesome properties. One such property is that there exists only a finite number of such floating point numbers. In fact, there are fewer than

$$10^4 \cdot 9$$

such numbers in $\bar{F}$, the largest such number in $\bar{F}$ being $.9999 \cdot 10^4$, while the smallest pcsitive number is $.0001 \cdot 10^{-4}$. Also, the spacing between two adjacent floating point numbers increases as the exponent becomes larger. That is, if

$$A = \left(\frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} + \frac{a_4}{10^4} \right) 10^e$$

and

$$A' = \left(\frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} + \frac{(a_4 - 1)}{10^4} \right) 10^e \; ,$$

then

$$d(A, A') = |A - A'|$$
$$= \left(\frac{1}{10^4}\right) 10^e$$
$$= 10^{e-4} .$$

Thus the collection $\bar{F}$ of floating point numbers does not satisfy Properties 1 and 2 of real numbers.

Another troublesome property of $\bar{F}$ is that real numbers do not have a unique representation in $\bar{F}$. For example, the real number .5 appears as four distinct elements of $\bar{F}$. They are

$$.5000 \cdot 10^0, \quad .0500 \cdot 10^1, \quad .0050 \cdot 10^2, \quad \text{and} \quad .0005 \cdot 10^3 .$$

An even more undesirable occurrence is that the real number zero corresponds to a total of nine elements of $\bar{F}$,

$$.0000 \cdot 10^{-4}, \quad .0000 \cdot 10^{-3}, \quad \ldots\ldots, \quad .0000 \cdot 10^3, \quad \text{and} \quad .0000 \cdot 10^4.$$

When two M-digit numbers are added, the result may not be an M-digit number. Therefore, in dealing with floating point numbers, a pseudo addition will be defined in such a manner that the pseudo sum of two M-digit numbers will again be an M-digit number. The pseudo operation will be indicated by an asterisk superscript showing truncation of the number of digits at M. An r subscript below the asterisk, when it appears, will indicate rounding has occurred by the addition of $\beta^{-M}/2$ (with the same sign as the term being rounded) prior to the truncation. The notation here is the same as that used by Householder [6] and Carr [3].

Let F be the set of all floating point numbers for given integers M, β, L, and U. Now floating point addition and multiplication will

be defined.

Definition 1.2: Psuedo Addition. Let $A, B \in F$ such that $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $e_2 \geq e_1$. The operation of pseudo addition, $\oplus$, is defined:

(i) If $|(a\beta^{e_1-e_2})_r^* + b| < 1$, then

$$A \oplus B = \left[(a\beta^{e_1-e_2})_r^* + b\right]\beta^{e_2} .$$

(ii) If $|(a\beta^{e_1-e_2})_r^* + b| \geq 1$ and $e_2 < U - 1$, then

$$A \oplus B = \left[(a\beta^{e_1-e_2-1})_r^* + (b\beta^{-1})_r^*\right]\beta^{e_2+1} .$$

(iii) If $|(a\beta^{e_1-e_2})_r^* + b| \geq 1$ and $e_2 = U - 1$, then

$A \oplus B$ is not defined.

Definition 1.3: Pseudo Multiplication. Let $A, B \in F$ such that $A = a\beta^{e_1}$ and $B = b\beta^{e_2}$. The operation of pseudo multiplication, $\otimes$, is defined:

(i) If $L < e_1 + e_2 < U$, then

$$A \otimes B = (a \cdot b)_r^* \beta^{e_1+e_2} .$$

(ii) If $e_1 + e_2 \leq L$ or $e_1 + e_2 \geq U$, then

$A \otimes B$ is not defined.

Definition 1.4: Floating Point Number System. Let F be the set of all floating point numbers for given integers M, β, L and U and let $\oplus$ and $\otimes$ be the operations as defined in Definitions 1.2 and 1.3 respectively. Then the system $\{F, \oplus, \otimes\}$ will be called a floating point number system.

Pseudo addition and pseudo multiplication are not defined for all elements of F since the operations may produce M-digit numbers whose exponents lie outside the bounded limits. Thus the floating point number system $\{F, \oplus, \otimes\}$ is not closed under pseudo addition or pseudo multiplication.

To further illustrate some of the difficulties in dealing with a floating point number system, consider the system $\{\bar{F}, \oplus, \otimes\}$, where again $\bar{F}$ will denote the set of all floating point numbers for $M = 4$, $\beta = 10$, $L = -5$, and $U = 5$. It was shown earlier that a real number could correspond to several elements in $\bar{F}$ and that each member of $\bar{F}$ was a representation of a real number.

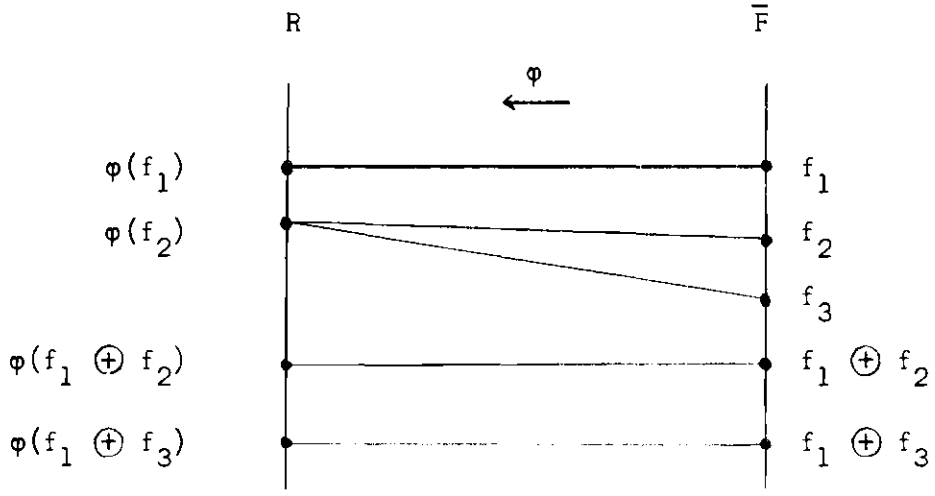

Figure 1. Illustration of Text.

If one adds different floating point representations of the same real number to another floating point number, then one could obtain floating point numbers which correspond to different real numbers. That is, if φ is the mapping of $\bar{F}$ into R so that a floating point number

maps into the real number which it represents, then

$$\varphi(f_1 \oplus f_2) \quad \text{and} \quad \varphi(f_1 \oplus f_3)$$

need not be equal even though

$$\varphi(f_2) = \varphi(f_3) .$$

For example, if $f_1 = .5000 \cdot 10^{-4}$, $f_2 = 0 \cdot 10^{4}$ and $f_3 = 0 \cdot 10^{-4}$, then

$$\begin{aligned} \varphi(f_1 \oplus f_2) &= \varphi(.5 \cdot 10^{-4} \oplus 0 \cdot 10^{4}) \\ &= \varphi\left(\left[(.5 \cdot 10^{-8})_r^* + 0\right] \cdot 10^{4}\right) \\ &= \varphi(0 + 0) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \varphi(f_1 \oplus f_3) &= \varphi(.5000 \cdot 10^{-4} \oplus 0 \cdot 10^{-4}) \\ &= \varphi\left(\left[.5000 + .0000\right] \cdot 10^{-4}\right) \\ &= \varphi(.5 \cdot 10^{-4}) \\ &= .5 \cdot 10^{-4} . \end{aligned}$$

It has been shown by means of examples that a floating point number system has many troublesome properties. Since $\{F, \oplus, \otimes\}$ is not closed under pseudo addition or pseudo multiplication, it cannot be a field. It is of interest, however, to determine just what properties the floating point number system does have.

Theorem 1.5: Let $A, B \in F$. If $A \oplus B \in F$, then $A \oplus B = B \oplus A$.

Proof. Obvious from Definition 1.3.

Theorem 1.6: Let $A, B \in F$. If $A \circledast B \in F$, then $A \circledast B = B \circledast A$.

Proof: Obvious from Definition 1.4.

Thus pseudo addition and pseudo multiplication are commutative operations when they are defined. However, the two operations are not associative as will be illustrated by the next two examples. For the next three examples $\overline{F}$ will again denote the set of all floating point numbers for $M = 4$, $\beta = 10$, $L = -5$, and $U = 5$.

Example 1.7: (Addition is non-associative). Let $A = .4004 \times 10^{0}$, $B = .3003 \times 10^{0}$, and $C = .8008 \times 10^{1}$.

Then

$$\begin{aligned} A \oplus (B \oplus C) &= .4004 \times 10^{0} \oplus (.3003 \times 10^{0} \oplus .8008 \times 10^{1}) \\ &= .4004 \times 10^{0} \oplus [(.03003)_{r}^{*} + .8008] \times 10^{1} \\ &= .4004 \times 10^{0} \oplus (.0300 + .8008) \times 10^{1} \\ &= .4004 \times 10^{0} + (.8308 \times 10^{1}) \\ &= [(.04004)_{r}^{*} + .8308] \times 10^{1} \\ &= (.0400 + .8308) \times 10^{1} \\ &= .8708 \times 10^{1} \end{aligned}$$

and

$$(A \oplus B) \oplus C = [.4004 \times 10^0 \oplus .3003 \times 10^0] \oplus (.8008) \times 10^1$$

$$= .7007 \times 10^0 \oplus (.8008) \times 10^1$$

$$= [(.07007)^*_r + .8008] \times 10^1$$

$$= (.0701 + .8008) \times 10^1$$

$$= .8709 \times 10^1 .$$

Thus $A \oplus (B \oplus C) \in \overline{F}$ and $(A \oplus B) \oplus C \in \overline{F}$ but $A \oplus (B \oplus C) \neq (A \oplus B) \oplus C$.

Example 1.8: (Multiplication is non-associative).

Let $A = .5004 \times 10^0$, $B = .4004 \times 10^0$, and $C = .8008 \times 10^0$.

Then

$$A \otimes (B \otimes C) = .5004 \times 10^0 \otimes (.4004 \times 10^0 \otimes .8008 \times 10^0)$$

$$= .5004 \times 10^0 \otimes (.32064032)^*_r \times 10^0$$

$$= .5004 \times 10^0 \otimes 3206 \times 10^0$$

$$= (.16042824)^*_r \times 10^0$$

$$= .1604 \times 10^0$$

and

$$(A \otimes B) \otimes C = (.5004 \times 10^0 \otimes .4004 \times 10^0) \otimes .8008 \times 10^0$$

$$= (.20036016)^*_r \times 10^0 \otimes .8008 \times 10^0$$

$$= .2004 \times 10^0 \otimes .8008 \times 10^0$$

$$= (.16048032)^*_r \times 10^0$$

$$= .1605 \times 10^0 .$$

Thus $(A \otimes B) \otimes C \in \overline{F}$ and $A \otimes (B \otimes C) \in \overline{F}$ but $(A \otimes B) \otimes C \neq A \otimes (B \otimes C)$.

The next example will show that the floating point number system does not possess a distributive law.

Example 1.9: Let $A = .9009 \times 10^0$, $B = .1234 \times 10^0$, and $C = .2345 \times 10^0$.

Then $A \otimes (B \oplus C) = .9009 \times 10^0 \otimes (.1234 \times 10^0 \oplus .2345 \times 10^0)$

$= .9009 \times 10^0 \otimes .3579 \times 10^0$

$= (.32243211)^*_r \times 10^0$

$= .3224 \times 10^0$

and

$(A \otimes B) \oplus (A \otimes C) = (.9009 \times 10^0 \otimes .1234 \times 10^0) \oplus$

$(.9009 \times 10^0 \otimes .2345 \times 10^0)$

$= (.11117106)^*_r \times 10^0 \oplus (.21126105)^*_r \times 10^0$

$= .1112 \times 10^0 \oplus .2113 \times 10^0$

$= .3225 \times 10^0$.

Hence $A \otimes (B \oplus C) \in \bar{F}$ and $(A \times B) + (A \times C) \in \bar{F}$, but $A \otimes (B \oplus C) \neq (A \otimes B) \oplus (A \otimes C)$.

It has been shown that the real number zero does not have a unique representation in floating point. However, the system $\{F, \oplus\}$ does have a unique identity, as will be shown in the next theorem.

Theorem 1.10: The system $\{F, \oplus\}$ has a unique identity element, $0 \cdot \beta^{L+1}$.

Proof:

Let $A = a\beta^e \in F$. Then $e \geq L + 1$ and

$$A \oplus 0 \cdot \beta^{L+1} = [(0 \cdot \beta^{L+1-e})^{*}_{r} + a]\,\beta^{e}$$

$$= (0 + a)\,\beta^{e}$$

$$= a\beta^{e}$$

$$= A\ .$$

Also,

$$0 \cdot \beta^{L+1} \oplus A = A \oplus 0 \cdot \beta^{L+1}$$

$$= A\ .$$

Thus $0 \cdot \beta^{L+1}$ is an identity for $\{F, \oplus\}$.

To show uniqueness, assume that there is an element $B \in F$ such that

$$A \oplus B = A \quad \text{for all} \quad A \in F\ .$$

Then

$$0 \cdot \beta^{L+1} \oplus B = 0 \cdot \beta^{L+1} \quad \text{since} \quad 0 \cdot \beta^{L+1} \in F$$

and

$$0 \cdot \beta^{L+1} \oplus B = B \quad \text{since} \quad 0 \cdot \beta^{L+1} \quad \text{is an identity for} \quad F.$$

Hence

$$B = 0 \cdot \beta^{L+1}\ .$$

Since zero does not have a unique representation in the floating point number system, each element of F does not have a pseudo additive inverse. For example, if $A = a\beta^{e}$, where $e > L + 1$, then

$$A \oplus (-A) = 0 \cdot \beta^{e} ,$$

but $0 \cdot \beta^{e}$ is not the additive identity, $0 \cdot \beta^{L+1}$.

The real number one also has many representations in the floating point number system. However, none of representations of one is an identity for $\{F, \otimes\}$ as is shown in the next example.

Example 1.11: Let $A = .4321 \cdot 10^{0}$ be a member of $\bar{F}$. Then

$$\begin{aligned} A \otimes (.1000) \cdot 10^{1} &= [(.4321)(.1000)]_{r}^{*} \cdot 10^{0+1} \\ &= (.04321)_{r}^{*} \cdot 10^{1} \\ &= (.0432) \cdot 10^{1} . \end{aligned}$$

Similarly,

$$A \otimes (.0100) \cdot 10^{2} = (.0043) \cdot 10^{2} ,$$

$$A \otimes (.0010) \cdot 10^{3} = (.0004) \cdot 10^{3} , \quad \text{and}$$

$$A \otimes (.0001) \cdot 10^{4} = (.0000) \cdot 10^{4} .$$

Thus none of the representations of one is an identity for $\bar{F}$.

In fact $\{F, \otimes\}$ does not have an identity as will be shown. Let $A = a\beta^{e_1}$. If $B = b\beta^{e_2}$ is an identity for $\{F, \otimes\}$, then

$$A \otimes B = A \quad \text{or}$$

$$(a \cdot b)_{r}^{*} \beta^{e_1+e_2} = a \cdot \beta^{e_1} .$$

Therefore, B must be such that

$$(a \cdot b)_{r}^{*} = a \quad \text{and}$$

$$e_2 = 0 .$$

Since $|b| < 1$, there is no $B = b\beta^0 \in F$ such that $(a \cdot b)_r^* = a$ for all $A \in F$. Hence, the system $\{F, \otimes\}$ does not have an identity.

Since a set of floating point numbers is a subset of the real numbers, it forms a metric space for the metric $d(x, y) = |x - y|$. However, if the function $\rho(x, y)$ were defined to be $\rho(x, y) = |x \oplus - y|$, then ρ is not a metric on F. This will be shown in the next example.

Example 1.12: (Lack of Triangle Law). Let X, Y, and $Z \in \bar{F}$ such that $X = .1005 \cdot 10^1$, $Y = -.1055 \cdot 10^0$, and $Z = -.1004 \cdot 10^0$. Then

$$\begin{aligned} |X \oplus - Y| &= |.1005 \cdot 10^1 \oplus .1055 \cdot 10^0| \\ &= [(.01055)_r^* + .1005] \cdot 10^1 \\ &= (.0106 + .1005) \cdot 10^1 \\ &= .1111 \cdot 10^1 , \end{aligned}$$

$$\begin{aligned} |X \oplus - Z| &= |.1005 \cdot 10 \oplus .1004 \cdot 10^0| \\ &= [(.01004)_r^* + .1005] \cdot 10^1 \\ &= (.0100 + .1005) \cdot 10^1 \\ &= .1105 \cdot 10^1 , \end{aligned}$$

and

$$\begin{aligned} |Z \oplus - Y| &= | - .1004 \cdot 10^0 \oplus .1055 \cdot 10^0| \\ &= (-.1004 + .1055) \cdot 10^0 \\ &= .0051 \cdot 10^0 . \end{aligned}$$

Then

$$|X \oplus -Z| \oplus |Z \oplus -Y| = (.1105 \cdot 10^{1}) \oplus (.0051 \cdot 10^{0})$$

$$= [(.00051)_{r}^{*} + .1105] \cdot 10^{1}$$

$$= (.0005 + .1105) \cdot 10^{1}$$

$$= .1110 \cdot 10^{1}$$

and

$$|X \oplus -Y| > |X \oplus -Z| \oplus |Z \oplus -Y| .$$

Thus $\{\overline{F}, \oplus\}$ does not have a triangle law.

To summarize some of the properties (or lack of properties) of the floating point number system, it was found that the collection of floating point numbers is a finite collection with the spacing between adjacent numbers increasing with the exponent. Some real numbers have many representations in the collection which produced some troublesome results. It was seen that the system was not closed under pseudo addition or pseudo multiplication. The pseudo operations were found to be commutative when defined, but not associative or distributive. Even though the system has a unique additive identity, there are some members of this floating point number system which do not have an additive inverse. This system has no multiplicative identity.

C. Normalized Floating Point Number System

A variation of the floating point number system of the preceding section is used in a number of digital computers. In this variation, all floating point numbers except zero have a nonzero leading digit in the fractional part. Special conventions are adopted to cover the case of zero. In the resulting system a maximum number of significant digits are

retained at each stage of calculation. This system will be called a normalized floating point number system. The normalized system will be defined and examined in this section.

Definition 1.13: Normalized Floating Point Number. A normalized floating point number is a floating point number $d \cdot \beta^e$ such that $\beta^{-1} \leq |d| < 1$ or is the floating point number $0 \cdot \beta^0$.

In this normalized system, after each pseudo operation which produces a zero in the first digit of the fractional part, the number is normalized or shifted until the first digit is different from zero. If the fractional part is zero, then the exponent is set to zero also. Since zero has to be considered as a special case in a normalized system, the representation of zero on some machines may differ from the one given here.

As an illustration, let $\bar{F}_N$ denote the set of all normalized floating point numbers for $M = 4$, $\beta = 10$, $L = -5$, and $U = 5$. Then any member of $\bar{F}_N$ other than zero is of the form

$$\pm\left(\frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3} + \frac{d_4}{10^4}\right) \cdot 10^e$$

where d_i $(i = 1,2,3,4)$ and e are integers which satisfy

$$1 \leq d_1 < 10$$

$$0 \leq d_i < 10 \quad i = 2,3,4$$

$$-5 < e < 5 .$$

Zero has the form $(0/10 + 0/10^2 + 0/10^3 + 0/10^4) \cdot 10^0$.

Note that $\bar{F}_N \subset \bar{F}$, and that the largest element of $\bar{F}_N$, $.9999 \cdot 10^4$, is the same as that of $\bar{F}$. However, the smallest positive element of $\bar{F}_N$, $.1000 \cdot 10^{-4}$, is larger than the smallest positive element of $\bar{F}$, $.0001 \cdot 10^{-4}$. Hence there are fewer members of $\bar{F}_N$ near zero than there were of $\bar{F}$.

It was seen that real numbers do not have a unique representation in $\bar{F}$. This is not the case for $\bar{F}_N$. If r is a four digit real number such that $10^{-5} < r < 10^5$, then r has a unique representation in $\bar{F}_N$.

It is interesting to note the spacing of the members of $\bar{F}_N$. Since $U = 5$ and $L = -5$, the number of members of $\bar{F}_N$ in the interval $(0, .1)$ is the same as the number in the interval $(1, \infty)$. Thus the distance between two members of $\bar{F}_N$ increases as the exponent increases, as was true for the set $\bar{F}$. However, the "relative spacing" of two adjacent members of $\bar{F}_N$ does not increase. That is, if

$$A = \left(\frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3} + \frac{a_4}{10^4}\right) 10^e \quad \text{and}$$

$$A' = \left(\frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^4} + \frac{(a_4-1)}{10^4}\right) 10^e,$$

then

$$\frac{|A - A'|}{10^e} = \frac{\left(\frac{1}{10^4}\right) \cdot 10^e}{10^e}$$

$$= 10^{-4}.$$

Thus the "relative spacing" $\dfrac{|A - A'|}{10^e}$ is a constant, 10^{-4}.

The pseudo operations for the normalized system are necessarily more involved than those of the non-normalized system since shifting of the fractional part has to be considered. However, when the numbers are in the normalized form and no shifting occurs, the definitions for the normalized and non-normalized operations are equivalent.

Let F_N be the set of all normalized floating point numbers for given integers M, β, L, and U. The pseudo operations of elements of F_N will now be defined.

Definition 1.14: Pseudo Addition. Let $A, B \in F_N$ such that $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $e_2 \geq e_1$. The operation of pseudo addition, $\oplus$, is defined for normalized floating point numbers as follows:

(i) If $\beta^{-1} \leq |(a\beta^{e_1-e_2})_r^* + b| < 1$, $a \neq 0$, and $b \neq 0$, then

$$A \oplus B = [(a\beta^{e_1-e_2})_r^* + b]\,\beta^{e_2}.$$

(ii) If $|(a\beta^{e_1-e_2})_r^* + b| \geq 1$ and $e_2 + 1 < U$, then

$$A \oplus B = [(a\beta^{e_1-e_2-1})_r^* + (b\beta^{-1})_r^*]\,\beta^{e_2+1}.$$

If $e_{2+1} \geq U$, $A \oplus B$ is not defined.

(iii) If $0 < |(a\beta^{e_1-e_2})_r^* + b| < \beta^{-1}$ and $e_2 + \lfloor \log_\beta |\sigma| \rfloor + 1 > L$,

$$A \oplus B = (\sigma\beta^{-\lfloor \log_\beta |\sigma| \rfloor - 1})\,\beta^{e_2 + \lfloor \log_\beta |\sigma| \rfloor + 1}$$

where the brackets indicate "the greatest integer less than or equal to" and $\sigma = (a\beta^{e_1-e_2})_r^* + b$.

If $e_2 + \lfloor \log_\beta |\sigma| \rfloor + 1 \leq L$, then $A \oplus B$ is not defined.

(iv) If $(a\beta^{e_1-e_2})_r^* + b = 0$, then

$$A \oplus B = 0 \cdot \beta^0 .$$

(v) If $A = 0 \cdot \beta^0$ then

$$A \oplus B = B .$$

Definition 1.15: Pseudo Multiplication. Let $A, B \in F_N$ such that $A = a\beta^{e_1}$ and $B = b\beta^{e_2}$. The operation of pseudo multiplication, $\otimes$, is defined for normalized floating point numbers as follows:

(i) If $A \neq 0$, $B \neq 0$, $e_1 + e_2 + \lfloor \log_\beta |a \cdot b| \rfloor + 1 < U$,

and $e_1 + e_2 + \lfloor \log_\beta |a \cdot b| \rfloor + 1 > L$,

then

$$A \otimes B = (a \cdot b\beta^{-\lfloor \log_\beta |a \cdot b| \rfloor - 1})_r^* \beta^{e_1+e_2+\lfloor \log_\beta |a \cdot b| \rfloor + 1} .$$

If $e_1 + e_2 + \lfloor \log_\beta |a \cdot b| \rfloor + 1 \geq U$ or $e_1 + e_2 + \lfloor \log_\beta |a \cdot b| \rfloor + 1 \leq L$, then $A \otimes B$ is not defined.

(ii) If $A = 0 \cdot \beta^0$ or $B = 0 \cdot \beta^0$, then

$$A \otimes B = 0 \cdot \beta^0 .$$

Definition 1.16: Normalized Floating Point Number System. Let F_N be a set of normalized floating point numbers and $\oplus$ and $\otimes$ be the operations defined in Definitions 1.14 and 1.15, respectively. The system $\{F_N, \oplus, \otimes\}$ will be called a normalized floating point number system.

Theorem 1.17: If $A \oplus B \in F_N$ and if $A \otimes B \in F_N$, then $A \oplus B = B \oplus A$

and $A \oplus B = B \oplus A$.

Proof: This is obvious from Definitions 1.14 and 1.15.

The systems $\{F_N, \oplus\}$ and $\{F_N, \otimes\}$ are not associative and $\{F_N, \oplus, \otimes\}$ does not possess a distributive law. Examples 1.7, 1.8, and 1.9, which apply to $\{F_N \oplus, \otimes\}$ since the numbers were in the normalized form and no shifting occurred, illustrate the lack of these properties.

Theorem 1.18: The system $\{F_N, \oplus\}$ has an identity element, $0 \cdot \beta^0$.

Proof: Let $A \in F_N$. Then by Definition 1.14,

$$A \oplus 0 \cdot \beta^0 = A .$$

Theorem 1.19: The identity of Theorem 1.18 is unique.

Proof: Assume that there is an element $A \in F_N$ such that

$$A \oplus B = B \quad \text{for all} \quad B \in F_N .$$

Then

$$A \oplus 0 \cdot \beta^0 = 0 \cdot \beta^0 \quad \text{since} \quad 0 \cdot \beta^0 \in F_N .$$

But

$$A \oplus 0 \cdot \beta^0 = A \quad \text{by Theorem 1.18.}$$

Hence

$$A = 0 \cdot \beta^0 .$$

Theorem 1.20: Every element of F_N has an additive inverse element in F_N.

Proof:

Let $A = a\beta^e$. If $A \in F_N$ then $-A \in F_N$ and

$$\begin{aligned} A \oplus (-A) &= a\beta^e \oplus -a\beta^e \\ &= [(a)_r^* + (-a)]\,\beta^e \\ &= [a + (-a)]\,\beta^e \\ &= 0 \cdot \beta^0 . \end{aligned}$$

The additive inverse is not necessarily unique, however. The next example shows that an element could have several inverses.

Example 1.21: Let $A, B \in \overline{F}_N$ such that $A = .9999 \cdot 10^0$ and $B = -.1000 \cdot 10^1$. Then

$$\begin{aligned} A \oplus B &= [(.09999)_r^* + (-.1000)] \cdot 10^1 \\ &= [(.09999 + .00005)^* + (-.1000)] \cdot 10^1 \\ &= [(.10004)^* + (-.1000)] \cdot 10^1 \\ &= [(.1000) + (-.1000)] \cdot 10^1 \\ &= 0 \cdot 10^1 \\ &= 0 \cdot 10^0 . \end{aligned}$$

Thus, $B \neq -A$ and

$$A \oplus B = 0 \cdot 10^0 .$$

Theorem 1.22: If $L < 1$, then the system $\{F_N, \otimes\}$ has a unique identity element, $\beta^{-1} \cdot \beta^1$.

Proof: Let $A \in F_N$ such that $A = a\beta^e$. If $A = 0 \cdot \beta^0$, then

$$A \otimes \beta^{-1} \cdot \beta^1 = \beta^{-1} \cdot \beta^1 \otimes A$$
$$= 0 \cdot \beta^0$$
$$= A .$$

If $a \neq 0$, then

$$A \otimes \beta^{-1} \cdot \beta^1 = \left((a \cdot \beta^{-1})\beta^{-\lfloor \log_\beta |a \cdot \beta^{-1}| \rfloor - 1}\right)_r^* \beta^{e+1+\lfloor \log_\beta |a\cdot\beta^{-1}| \rfloor + 1}$$
$$= \left(a \cdot \beta^{-1}\, \beta^{-\lfloor -2 \rfloor - 1}\right)_r^* \beta^{e+1+\lfloor -2 \rfloor + 1}$$
$$= (a)_r^* \beta^e$$
$$= a\beta^e$$
$$= A .$$

Also

$$\beta^{-1} \cdot \beta^1 \otimes A = A \otimes \beta^{-1} \cdot \beta^1$$
$$= A .$$

Hence

$$A \otimes \beta^{-1} \cdot \beta^1 = \beta^{-1} \cdot \beta^1 \otimes A$$
$$= A ,$$

and $\beta^{-1} \cdot \beta$ is an identity for $\{F_N, \otimes\}$.

To show uniqueness, assume that there is an element $B \in F_N$ such that

$$A \otimes B = A \quad \text{for all} \quad A \in F_N .$$

Then

$$\beta^{-1} \cdot \beta^{1} \otimes B = \beta^{-1} \cdot \beta^{1} \quad \text{since} \quad \beta^{-1} \cdot \beta^{1} \in F_N$$

and

$$\beta^{-1} \cdot \beta^{1} \otimes B = B \quad \text{since} \quad \beta^{-1} \cdot \beta^{1} \quad \text{is an identity.}$$

Hence

$$B = \beta^{-1} \cdot \beta^{1} .$$

Even though the system $\{F_N, \otimes\}$ has an identity, some elements of F_N do not have a multiplicative inverse as will be shown.

Example 1.23: (Lack of Multiplicative Inverse).

Let $A = .9999 \cdot 10^{0}$, $B = .1000 \cdot 10^{1}$, and $C = .1001 \cdot 10^{1}$, where A, B, and $C \in \bar{F}_N$. Then

$$\begin{aligned} A \otimes B &= (.9999 \cdot 10^{0}) \otimes (.1000 \cdot 10^{1}) \\ &= .9999 \cdot 10^{0} \end{aligned}$$

and

$$\begin{aligned} A \otimes C &= (.9999 \cdot 10^{0}) \otimes (.1001 \cdot 10^{1}) \\ &= (.10008999)^{*}_{r} \cdot 10^{1} \\ &= .1001 \cdot 10^{1} . \end{aligned}$$

Since $\bar{F}_N$ does not possess Property 1 of the real numbers, there is no member of $\bar{F}_N$ between $.1000 \cdot 10^{1}$ and $.1001 \cdot 10^{1}$. Thus there is no element $\bar{A} \in \bar{F}_N$ such that

$$A \otimes \bar{A} = .1000 \cdot 10^{1} .$$

As was true with the floating point number system, the normalized system does not possess a triangle law. Example 1.12 applied to $\{F_N, \oplus, \otimes\}$ since the numbers were in the normalized form and no shifting occurred.

Like the floating point number system, the normalizing floating point number system does not have distributive, associative, or triangular laws. However, unlike the floating point number system, two distinct normalized floating point numbers cannot represent the same real number. There is a constant "relative" spacing between these numbers. The normalized system possesses unique additive and multiplicative identities. Each element has an additive inverse but the inverse is not necessarily unique. However some elements do not have a multiplicative inverse. The pseudo operations again were found to be commutative when defined.

CHAPTER II

PSEUDO ARITHMETIC OPERATIONS

In Chapter I, a floating point number system and a normalized floating point number system were defined and the algebraic properties of each examined. It was seen that the normalized system had more properties of the real number system than did the non-normalized system. The normalized system also has the property that a maximum number of significant digits are retained at each stage of computation. However, the normalizing procedure could introduce meaningless digits into the numbers and the number of significant digits is not known at the end of computation. For this reason, there has been increasing interest in the floating point number system. Carr [3] discusses this "significant" system and derives error bounds for the pseudo operations. Ashenhurst and Metropolis [1] describe a variation of the system wherein the numbers are not normalized except where absolutely necessary.

Since the normalized floating point number system is most often used in digital computers, the remainder of this study will be devoted to the normalized system. In this chapter, the pseudo operations of the normalized floating point number system will be compared with the corresponding real number operations. The purpose of the comparison is to establish bounds on the error created by performing pseudo operations in lieu of real number operations.

Theorem 2.1: Let $A, B \in F_N$ such that $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $e_2 \geq e_1$.

A bound on the error generated by performing pseudo addition instead of real number addition is

$$|(A \oplus B) - (A + B)| \leq 2\varepsilon \beta^{e_2+1} , \tag{2.1}$$

where $\varepsilon = \beta^{-M}/2$.

Proof:

Case 1. $$\beta^{-1} \leq |(a\beta^{e_1-e_2})^*_r + b| < 1$$

$$A \oplus B = \left[(a\beta^{e_1-e_2})^*_r + b \right] \beta^{e_2} .$$

Then

$$A \oplus B = \left[(a\beta^{e_1-e_2} + \eta) + b \right] \beta^{e_2}$$

where $-\varepsilon \leq \eta < \varepsilon$, ε being the maximum roundoff in the $(M + 1)^{st}$ digit of the fractional part. For pseudo addition as defined in Chapter I, $\varepsilon = \beta^{-M}/2$. If pseudo addition had been defined in such a manner that the fractional part was truncated after M digits without altering the M^{th} digit itself, then $\varepsilon = \beta^{-M}$. This latter method is used on some computing machines.

$$(A \oplus B) - (A + B) = \left[(a\beta^{e_1-e_2} + \eta) + b \right] \beta^{e_2} - \left[a\beta^{e_1-e_2} + b \right] \beta^{e_2}$$

$$= \eta \beta^{e_2} .$$

Therefore,

$$|(A \oplus B) - (A + B)| \leq \varepsilon \beta^{e_2} . \tag{2.2}$$

Case 2. $|(a\beta^{e_1-e_2})_r^* + b| \geq 1$

$$A \oplus B = \left[(a\beta^{e_1-e_2-1})_r^* + (b\beta^{-1})_r^* \right] \beta^{e_2+1}$$

$$= \left[a\beta^{e_1-e_2-1} + \eta_1 + b\beta^{-1} + \eta_2 \right] \beta^{e_2+1} ,$$

where

$$-\varepsilon \leq \eta_1 < \varepsilon$$

$$-\varepsilon \leq \eta_2 < \varepsilon .$$

$$(A \oplus B) - (A + B) = (a\beta^{e_1-e_2-1} + \eta_1 + b\beta^{-1} + \eta_2)\beta^{e_2+1} - (a\beta^{e_1-e_2-1} + b\beta^{-1})\beta^{e_2+1}$$

$$= (\eta_1 + \eta_2)\beta^{e_2+1} .$$

Therefore,

$$|(A \oplus B) - (A + B)| \leq 2\varepsilon\beta^{e_2+1} . \qquad (2.3)$$

Case 3. $0 < |(a\beta^{e_1-e_2})_r^* + b| < \beta^{-1}$

$$A \oplus B = \sigma\beta^{-\lfloor \log_\beta |\sigma| \rfloor - 1}\beta^{e_2+\lfloor \log_\beta |\sigma| \rfloor + 1} ,$$

where $\sigma = (a\beta^{e_1-e_2})_r^* + b .$

Then

$$(A \oplus B) - (A + B) = \left[(a\beta^{e_1-e_2})^*_r + b \right]\beta^{-\lfloor \log_\beta |\sigma| \rfloor - 1}\ \beta^{e_2+\lfloor \log_\beta |\sigma| \rfloor + 1}$$

$$- (a\beta^{e_1-e_2} + b)\beta^{e_2}$$

$$(A \oplus B) - (A + B) = \left[(a\beta^{e_1-e_2} + \eta) + b \right]\beta^{e_2} - (a\beta^{e_1-e_2} + b)\beta^{e_2}$$

$$= \eta\, \beta^{e_2} .$$

Therefore,

$$|(A \oplus B) - (A + B)| \leq \varepsilon \beta^{e_2} . \qquad (2.4)$$

Case 4. $$(a\beta^{e_1-e_2})^*_r + b = 0$$

$$A \oplus B = 0$$

$$= A + B$$

and

$$|(A \oplus B) - (A + B)| = 0 .$$

Therefore, the maximum difference occurs in Case 2. That is

$$|(A \oplus B) - (A + B)| \leq 2\varepsilon\beta^{e_2+1} .$$

Theorem 2.2: The error bound of Theorem 2.1 is the smallest bound that can be obtained for pseudo addition.

Proof: Let $A = [1 \cdot \beta^{-1} + (\beta/2)\beta^{-M}]\beta^{e}$ and $B = [(\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}]\beta^{e}$. Then

$$A + B = [1 \cdot \beta^{-1} + (\beta/2)\beta^{-M} + (\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}]\beta^{e}$$

$$= (\beta\beta^{-1} + \beta\beta^{-M})\beta^{e}$$

$$= (\beta^{-1} + \beta^{-M})\beta^{e+1} ,$$

and

$$A \oplus B = [(\beta^{-2} + (\beta/2)\beta^{-1-M})_{r}^{*} + ((\beta-1)\beta^{-2} + (\beta/2)\beta^{-1-M})_{r}^{*}]\beta^{e+1}$$

$$= [(\beta^{-2} + (\beta/2)\beta^{-1-M} + \beta^{-M}/2)^{*} + ((\beta-1)\beta^{-2} + (\beta/2)\beta^{-1-M} +$$

$$+ \beta^{-M}/2)^{*}]\beta^{e+1}$$

$$= (\beta^{-2} + \beta^{-M} + (\beta-1)\beta^{-2} + \beta^{-M})\beta^{e+1}$$

$$= (\beta\beta^{-2} + 2\beta^{-M})\beta^{e+1}$$

$$= (\beta^{-1} + 2\beta^{-M})\beta^{e+1} .$$

Therefore,

$$|(A \oplus B) - (A + B)| = |(\beta^{-1} + 2\beta^{-M})\beta^{e+1} - (\beta^{-1} + \beta^{-M})\beta^{e+1}|$$

$$= \beta^{-M}\beta^{e+1} .$$

But $\varepsilon = \beta^{-M}/2$, so that

$$|(A \oplus B) - (A + B) = 2\varepsilon\beta^{e+1} .$$

Thus the bound is actually obtained in this case.

Theorem 2.3: Let $A, B \in F_N$ such that $A = a\beta^{e_1}$ and $B = b\beta^{e_2}$. A bound on the error generated in pseudo multiplication is

$$|(A \circledast B) - (A \times B)| \leq \varepsilon \beta^{e_1+e_2} . \qquad (2.5)$$

Proof:

Case 1. $A \neq 0$ and $B \neq 0$.

$$A \circledast B = (ab\beta^{-\lfloor \log_\beta |ab| \rfloor - 1})_r \beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1}$$

$$= (ab\beta^{-\lfloor \log_\beta |ab| \rfloor - 1} + \eta)\beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1},$$

where $-\varepsilon \leq \eta < \varepsilon$. Then

$$(A \circledast B) - (A \times B) = (ab\beta^{-\lfloor \log_\beta |ab| \rfloor - 1} + \eta)\beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1} - ab\beta^{e_1+e_2}$$

$$(A \circledast B) - (A \times B) = \eta \beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1} .$$

Then

$$|(A \circledast B) - (A \times B)| \leq \varepsilon \beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1} . \qquad (2.6)$$

Here $e_1 + e_2 + \lfloor \log_\beta |ab| \rfloor + 1$ is the exponent of the final result; so an error bound can be stated in terms of the final product.

Since $\beta^{-1} \leq |a| < 1$ and $\beta^{-1} \leq |b| < 1$,

$$\beta^{-2} \leq |ab| < 1 \quad \text{and} \quad -2 \leq \log_\beta |ab| < 0 .$$

So

$$-2 \leq \lfloor \log_\beta |ab| \rfloor \leq -1 \quad \text{and}$$

$$\beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1} \leq \beta^{e_1+e_2} .$$

$$|(A \otimes B) - (A \times B)| \leq \varepsilon \beta^{e_1+e_2+\lfloor \log_\beta |ab| \rfloor + 1}$$

$$|(A \otimes B) - (A \times B)| \leq \varepsilon \beta^{e_1+e_2} .$$

Case 2. $A = 0$ or $B = 0$

$$A \otimes B = 0$$

$$= A \times B$$

and

$$|(A \otimes B) - (A \times B)| = 0.$$

Hence a bound for the error will be that of Case 1,

$$|(A \otimes B) - (A \times B)| \leq \varepsilon \beta^{e_1+e_2} .$$

Theorem 2.4: The error bound of Theorem 2.3 is the smallest bound that can be obtained for pseudo multiplication.

Proof: Let $A = [(\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}]\beta^{e_1}$ and $B = [(\beta-1)\beta^{-1} + 0\beta^{-M}]\beta^{e_2}$.

$$A \otimes B = \left\{ [(\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}][(\beta-1)\beta^{-1}]\beta^{-\lfloor \log_\beta |[(\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}][(\beta-1)\beta^{-1}]| \rfloor - 1} \right\}_r^* \times$$

$$\beta^{e_1+e_2+\lfloor \log_\beta |[(\beta-1)\beta^{-1} + (\beta/2)\beta^{-M}][(\beta-1)\beta^{-1}]| \rfloor + 1}$$

$$A \circledast B = \left[(1-2\beta^{-1}+2\beta^{-2}+(\beta/2)(\beta-1)\beta^{-1-M})\beta^{-(-1)-1}+\beta^{-M}/2\right]^{*} \beta^{e_1+e_2-1+1}$$

$$= \left[(\beta-2)\beta^{-1}+2\beta^{-2}+(\beta-1)/2\ \beta^{-M}+\beta^{-M}/2\right]\beta^{e_1+e_2}$$

$$= \left[(\beta-2)\beta^{-1}+2\beta^{-2}+\beta/2\ \beta^{-M}\right]\beta^{e_1+e_2}\ .$$

$$A \times B = [(\beta-1)\beta^{-1}+(\beta/2)\beta^{-M}][(\beta-1)\beta^{-1}]\beta^{e_1+e_2}$$

$$= [(\beta-2)\beta^{-1}+2\beta^{-2}+(\beta-1)/2\ \beta^{-M}]\ \beta^{e_1+e_2}\ .$$

Then

$$(A \circledast B) - (A \times B) = [(\beta-2)\beta^{-1}+2\beta^{-2}+(\beta/2)\beta^{-M}]\beta^{e_1+e_2}$$

$$- [(\beta-2)\beta^{-1}+2\beta^{-2}+(\beta-1)/2\ \beta^{-M}]\ \beta^{e_1+e_2}$$

$$(A \circledast B) - (A \times B) = [(\beta/2-(\beta-1)/2)]\beta^{-M}\beta^{e_1+e_2}$$

$$= (\beta^{-M}/2)\beta^{e_1+e_2}\ .$$

$$|(A \circledast B) - (A \times B)| = (\beta^{-M}/2)\beta^{e_1+e_2}$$

$$= \varepsilon\beta^{e_1+e_2}\ .$$

Hence the bound is obtained in this particular case.

In the real number system, the operations of subtraction and division are defined in terms of the basic operations, addition and multiplication, and the inverses of the elements involved. It was shown in Chapter I that an element of F_N could have several additive inverses. Therefore, it is necessary to define pseudo subtraction.

Definition 2.5: Pseudo Subtraction. Let $A, B \in F_N$. If $A \oplus -B \in F_N$, then pseudo subtraction, $\ominus$, is defined

$$A \ominus B = A \oplus -B .$$

If $A \oplus -B \notin F_N$, then $A \ominus B$ is not defined.

It was shown in Chapter I that $A \oplus -A = 0 \cdot \beta^0$. Thus

$$A \ominus A = A \oplus -A$$

$$= 0 \cdot \beta^0 ,$$

and a number pseudo subtracted from itself gives the additive identity, $0 \cdot \beta^0$.

Since pseudo subtraction is defined in terms of pseudo addition, the error bounds of Theorem 2.1 apply to subtraction.

It was shown in Chapter I that some members of F_N do not have a multiplicative inverse. Therefore pseudo division will have to be defined as a special operation. The definition given will be for a computer with a single length accumulator to hold the quotient. Therefore, it would not have available an extra position in which to add in $\beta^{-M}/2$ before the truncation.

Definition 2.6: Pseudo Division. Let $A, B \in F_N$ such that $A = a\beta^{e_1}$, $B = b \cdot \beta^{e_2}$, and $B \neq 0 \cdot \beta^0$. Pseudo division, $\oslash$, is defined:

(i) If $\beta^{-1} \leq |(a/b)^*| < 1$ and $e_1 - e_2 < U$

and $e_1 - e_2 > L$, then

$$A \oslash B = (a/b)^* \beta^{e_1 - e_2} .$$

If $e_1 - e_2 \geq U$ or $e_1 - e_2 \leq L$, then $A \oslash B$ is not defined.

(ii) If $0 < |(a/b)^*| < \beta^{-1}$ and $e_1 - e_2 + \lfloor \log_\beta |(a/b)^*| \rfloor + 1 < U$ and $> L$, then

$$A \oslash B = \left\{ (a/b)^* \ \beta^{-\lfloor \log_\beta |(a/b)^*| \rfloor - 1} \right\} \beta^{e_1 - e_2 + \lfloor \log_\beta |(a/b)^*| \rfloor + 1} .$$

If $e_1 - e_2 + \lfloor \log_\beta |(a/b)^*| \rfloor + 1 \geq U$ or $\leq L$, then $A \oslash B$ is undefined.

(iii) If $|(a/b)^*| > 1$ and $e_1 - e_2 + 1 < U$ and $> L$, then

$$A \oslash B = \left(\frac{(a\beta^{-1})^*_r}{b} \right)^* \beta^{e_1 - e_2 + 1} .$$

If $e_1 - e_2 + 1 \geq U$ or $\leq L$, then $A \oslash B$ is undefined.

(iv) If $(a/b)^* = 0$, then

$$A \oslash B = 0 \cdot \beta^0 .$$

Theorem 2.7: Let $A, B \in F_N$ such that $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $b \neq 0$. A bound for the error in pseudo division is

$$|(A \oslash B) - (A/B)| < (\beta + 2)\varepsilon \beta^{e_1 - e_2 + 1} . \qquad (2.7)$$

Proof:

Case 1. $\beta^{-1} \leq |(a/b)^*| < 1$

$$A \oslash B = (a/b)^* \beta^{e_1 - e_2}$$

$$= (a/b + \eta)\beta^{e_1 - e_2} ,$$

where $-2\varepsilon < \eta < 2\varepsilon$.

$$(A \oslash B) - (A/B) = (a/b + \eta)\beta^{e_1 - e_2} - a/b \ \beta^{e_1 - e_2}$$

$$= \eta \ \beta^{e_1 - e_2}$$

and

$$|(A \oslash B) - A/B| < 2\varepsilon\,\beta^{e_1-e_2}.$$

Case 2. $$0 < |(a/b)^*| < \beta^{-1}$$

$$A \oslash B = \left\{(a/b)^*\ \beta^{-\lfloor \log_\beta |(a/b)^*| \rfloor - 1}\right\} \beta^{e_1-e_2+\lfloor \log_\beta |(a/b)^*| \rfloor + 1}$$

$$= \left\{(a/b + \eta)\beta^{-\lfloor \log_\beta |(a/b)^*| \rfloor - 1}\right\} \beta^{e_1-e_2+\lfloor \log_\beta |(a/b)^*| \rfloor + 1},$$

where $-2\varepsilon < \eta < 2\varepsilon$.

$$(A \oslash B) - (A/B) = (a/b + \eta)\beta^{e_1-e_2} - a/b\ \beta^{e_1-e_2}$$

$$= \eta\,\beta^{e_1-e_2}.$$

$$|(A \oslash B) + (A/B)| < 2\varepsilon\,\beta^{e_1-e_2}.$$

Case 3. $$|(a/b)^*| > 1$$

$$A \oslash B = \left(\frac{(a\beta^{-1})^*_r}{b}\right)^*\ \beta^{e_1-e_2+1},$$

$$= \left(\frac{a\beta^{-1} + \eta_1}{b} + \eta_2\right) \beta^{e_1-e_2+1},$$

where $-\varepsilon \le \eta_1 < \varepsilon$

$-2\varepsilon < \eta_2 < 2\varepsilon$.

Then

$$(A \oslash B) - (A/B) = \left(\frac{a\beta^{-1} + \eta_1}{b} + \eta_2\right)\beta^{e_1-e_2+1} - a/b\ \beta^{e_1-e_2}$$

$$= (\eta_1/b + \eta_2)\ \beta^{e_1-e_2+1} ,$$

and

$$|(A \oslash B) - (A/B)| < (\varepsilon/|b| + 2\varepsilon)\beta^{e_1-e_2+1}$$

$$< (\varepsilon/\beta^{-1} + 2\varepsilon)\beta^{e_1-e_2+1}$$

$$< (\varepsilon\beta + 2\varepsilon)\beta^{e_1-e_2+1}$$

$$|(A \oslash B) - (A/B)| < (\beta+2)\varepsilon\beta^{e_1-e_2+1} .$$

Case 4. $(a/b)^* = 0$

Since $|b| > \beta^{-1}$, $(a/b)^* = 0$ only if $a = 0$ and

$$|(A \oslash B) - (A/B)| = 0 - 0$$

$$= 0 .$$

Since $(\beta+2)\,\varepsilon\,\beta^{e_i-e_2+1} > 2\varepsilon\,\beta^{e_1-e_2}$ for all e_1 and e_2, the bound for pseudo division would be that of Case 3,

$$|(A \oslash B) - (A/B)| < (\beta+2)\,\varepsilon\,\beta^{e_1-e_2+1} .$$

In summary, the error bounds for the pseudo operations are:

$$|(A \oplus B) - (A+B)| \leq 2\varepsilon\,\beta^{e_2+1} ,$$

$$|(A \otimes B) - (A\times B)| \leq \varepsilon\ \beta^{e_1+e_2} , \text{ and}$$

$$|(A \oslash B) - (A/B)| \leq (\beta+2)\,\varepsilon\,\beta^{e_1-e_2+1} ,$$

where $\varepsilon = \beta^{-M}/2$.

CHAPTER III

ERRORS DUE TO THE ABSENCE OF ASSOCIATIVE AND DISTRIBUTIVE LAWS

It was shown in Chapter I that pseudo addition and pseudo multiplication are not associative. Therefore, the order in which a sequence of pseudo operations is performed can make a difference in the generated error of the operations. In this chapter several sequences of operations will be examined to determine which order will produce the smallest error bound. Again the results will be for a normalized floating point number system.

Theorem 3.1: Let $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, $C = c\beta^{e_3}$, and $e_1 < e_2 < e_3$. The smallest error bound is obtained for the pseudo addition of A, B, and C by performing the operations in the order $(A \oplus B) \oplus C$.

Proof:

$$|(A \oplus B) \oplus C - (A+B+C)| = |(A \oplus B) \oplus C - [(A \oplus B) + C] + [(A \oplus B) + C] - (A+B+C)|$$

$$|(A \oplus B) \oplus C - (A+B+C)| \leq |(A \oplus B) \oplus C - [(A \oplus B) + C]| + |(A \oplus B) + C - (A+B+C)|.$$

If $A \oplus B = d\beta^{e}$, then $e \leq e_2 + 1 \leq e_3$ and by Equation 2.1,

$$|(A \oplus B) \oplus C - [(A \oplus B) + C]| \leq 2\varepsilon\beta^{e_3+1}$$

and

$$|[(A \oplus B) - C] - (A+B+C)| = |(A \oplus B) - (A+B)|$$
$$\leq 2\varepsilon\, \beta^{e_2+1} .$$

So

$$|(A \oplus B) \oplus C - (A+B+C)| \leq 2\varepsilon\, \beta^{e_3+1} + 2\varepsilon\, \beta^{e_2+1}$$
$$\leq 2\varepsilon\, (\beta^{e_3+1} + \beta^{e_2+1})$$
$$\leq 2\varepsilon\, \beta(\beta^{e_3} + \beta^{e_2}) .$$

In a similar manner,

$$|(A \oplus C) \oplus B - (A+C+B)| \leq |(A \oplus C) \oplus B - [(A \oplus C) + B]|$$
$$+ |(A \oplus C) + B - (A+B+C)| .$$

If $A \oplus C = e\beta^{e'}$, then $e' \leq e_3 + 1$ and by Equation 2.1,

$$|(A \oplus C) \oplus B - [(A \oplus C) + B]| \leq 2\varepsilon\, \beta^{e_3+2}$$

and

$$|(A \oplus C) + B - (A+B+C)| = |(A \oplus C) - (A+C)|$$
$$\leq 2\varepsilon\, \beta^{e_3+1} .$$

So

$$|(A \oplus C) \oplus B - (A+B+C)| \leq 2\varepsilon\, \beta^{e_3+2} + 2\varepsilon\, \beta^{e_3+1}$$
$$\leq 2\varepsilon\, \beta^{e_3+1}\, (\beta + 1) .$$

Also,

$$|(B \oplus C) \oplus A - (B+C+A)| \leq |(B \oplus C) \oplus A - [(B \oplus C) + A]| + |(B \oplus C) + A - (B+C+A)| .$$

If $B \oplus C = f\beta^{e''}$, then $e'' \leq e_3 + 1$ and

$$|(B \oplus C) \oplus A - [(B \oplus C) + A]| \leq 2\varepsilon\beta^{e_3+2}$$

and

$$|(B \oplus C) - (B+C)| \leq 2\varepsilon\beta^{e_3+2} .$$

So

$$|(B \oplus C) \oplus A - (B+C+A)| \leq 2\varepsilon\beta^{e_3+2} + 2\varepsilon\beta^{e_3+1} \leq 2\varepsilon\beta^{e_3+1}(\beta+1) .$$

Since $e_3 > e_2$, $2\varepsilon\beta(\beta^{e_3} + \beta^{e_2}) < 2\varepsilon\beta^{e_3+1}(\beta+1)$ and the smallest bound is $2\varepsilon\beta(\beta^{e_3} + \beta^{e_2})$, which is the bound for $|(A \oplus B) + C - (A+B+C)|$.

Thus the error bound for the pseudo addition of three numbers depends on the order of summation. It is smallest if the two numbers with the smallest exponents are added first. Wilkinson [14] shows that the upper bound for the error in summing a series of numbers is smallest if the terms are added in order of increasing absolute magnitude. Although the upper bound is smallest, the order does not necessarily give the smallest error. However, in determining a policy to follow in adding a series of numbers, the best policy would be to add in order of increasing absolute magnitude.

Theorem 3.2: Let $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $C = c\beta^{e_3}$. Suppose further that $|a| \leq |b| \leq |c|$ and $(B \otimes C) \otimes A = d_1\beta^{e_A}$, $(A \otimes C) \otimes B = d_2\beta^{e_B}$ and $(A \otimes B) \otimes C = d_3\beta^{e_C}$. If $e_A = e_B = e_C$, then the smallest error bound is obtained for the pseudo multiplication of A, B, and C by performing the operations in the order $(B \otimes C) \otimes A$.

Proof:

$$|(A \otimes B) \otimes C - A \times B \times C| \leq |(A \otimes B) \otimes C - (A \otimes B) \times C| + |(A \otimes B) \times C - (A \times B \times C)|.$$

By Equation 2.6, $|(A \otimes B) \otimes C - (A \otimes B) \times C| \leq \varepsilon\beta^{e_4}$, where e_4 is the exponent of the final product $(A \otimes B) \otimes C$.

$$|(A \otimes B) \times C - (A \times B \times C)| = |C|\,|(A \otimes B) - (A \times B)| \leq |C|\,\varepsilon\,\beta^{e_1+e_2}.$$

Therefore,

$$|(A \otimes B) \otimes C - (A \times B \times C)| \leq \varepsilon\,\beta^{e_4} + |C|\,\varepsilon\,\beta^{e_1+e_2}.$$

Similarly

$$|(A \otimes C) \otimes B - (A \times B \times C)| \leq \varepsilon\,\beta^{e'_4} + |B|\,\beta^{e_2+e_3}$$

and

$$|(B \otimes C) \otimes A - (A \times B \times C)| \leq \varepsilon\,\beta^{e''_4} + |A|\,\beta^{e_2+e_3},$$

where e'_4 is the exponent of $(A \otimes C) \otimes B$ and e''_4 is the exponent of $(B \otimes C) \otimes A$. Since it is assumed that $e_4 = e'_4 = e''_4$,

$$|(A \circledast B) \circledast C - (A \times B \times C)| \leq \varepsilon \beta^{e_4} + |C| \varepsilon \beta^{e_1+e_2},$$

$$|(A \circledast C) \circledast B - (A \times B \times C)| \leq \varepsilon \beta^{e_4} + |B| \varepsilon \beta^{e_1+e_3},$$

and

$$|(B \circledast C) \circledast A - (A \times B \times C)| \leq \varepsilon \beta^{e_4} + |A| \varepsilon \beta^{e_2+e_3}.$$

$$|C| \varepsilon \beta^{e_1+e_2} = (|c| \beta^{e_3}) \varepsilon \beta^{e_1+e_2}$$

$$= \varepsilon \beta^{e_1+e_2+e_3} |c| .$$

$$|B| \varepsilon \beta^{e_1+e_3} = (|b| \beta^{e_2}) \varepsilon \beta^{e_1+e_3}$$

$$= \varepsilon \beta^{e_1+e_2+e_3} |b| .$$

$$|A| \varepsilon \beta^{e_2+e_3} = (|a| \beta^{e_1}) \varepsilon \beta^{e_2+e_3}$$

$$= \varepsilon \beta^{e_1+e_2+e_3} |a| .$$

Since it is assumed that $|a| \leq |b| \leq |c|$,

$$\varepsilon\beta^{e_4} + |a|\varepsilon\beta^{e_1+e_2+e_3} \leq \varepsilon\beta^{e_4} + |b|\varepsilon\beta^{e_1+e_2+e_2} \leq \varepsilon\beta^{e_4} + |c|\varepsilon\beta^{e_1+e_2+e_3}.$$

Hence the bound obtained by forming the product $(B \circledast C) \circledast A$, $\varepsilon\beta^{e_4} + |a|\varepsilon\beta^{e_1+e_2+e_3}$, is the smallest.

In pseudo multiplying three floating point numbers, the smallest error bound is obtained if the two with the largest fractional parts are multiplied first. In multiplying a series of numbers, it would not be known in advance what the fractional parts of the partial products will be. Therefore, no general method could be stated. However, it appears

that it might be well to arrange the numbers in order of the magnitude of the fractional parts and to multiply them in this order.

The next problem to be encountered is the absence of a distributive law in the normalized floating point number system.

Theorem 3.3: Let $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $C = c\beta^{e_3}$, with $e_3 \geq e_2$. If $B \oplus C = d\beta^{e_4}$, an error bound for the sequence of operations $A \otimes (B \oplus C)$ is $\varepsilon\beta^{e_1+e_4} + |A|2\varepsilon\beta^{e_3+1}$.

Proof:

$$|A \otimes (B \oplus C) - A \times (B+C)| \leq |A \otimes (B \oplus C) - A \times (B \oplus C)| + |A \times (B \oplus C) - A \times (B+C)| .$$

By Equation 2.5, $|A \otimes (B \oplus C) - A \times (B \oplus C)| \leq \varepsilon\beta^{e_1+e_4}$.

$$|A \times (B \oplus C) - A \times (B+C)| = |A|\,|B \oplus C - (B+C)|$$

and by Equation 2.1, $|(B \oplus C) - (B+C)| \leq 2\varepsilon\beta^{e_3+1}$. Hence

$$|A \otimes (B \oplus C) - A \times (B+C)| \leq \varepsilon\beta^{e_1+e_4} + |A|2\varepsilon\beta^{e_3+1}.$$

Theorem 3.4: Let $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $C = c\beta^{e_3}$, with $e_3 \geq e_2$. An error bound for the sequence of operations $(A \otimes B) \oplus (A \otimes C)$ is $\varepsilon\beta^{e_1}(2\beta^{e_3+1} + \beta^{e_2} + \beta^{e_3})$.

Proof:

$$|(A \otimes B) \oplus (A \otimes C) - [(A \times B) + (A \times C)]| \leq |(A \otimes B) \oplus (A \otimes C) - [(A \otimes B) + (A \otimes C)]| + |(A \otimes B) - (A \times B)| + |(A \otimes C) - (A \times C)| .$$

If $A \otimes B = d\beta^{e}$ and $A \otimes C = f\beta^{e'}$, then $e \leq e_1 + e_2$ and $e' \leq e_1 + e_3$. Therefore by Equation 2.1,

$$|(A \otimes B) \oplus (A \otimes C) - [(A \otimes B) + (A \otimes C)]| \leq 2\varepsilon\beta^{e_1+e_3+1}.$$

By Equation 2.5, $|(A \otimes B) - (A \times B)| \leq \varepsilon\beta^{e_1+e_2}$ and

$$|(A \otimes C) - (A \times C)| \leq \varepsilon\beta^{e_1+e_3}.$$

Therefore,

$$|(A \otimes B) \oplus (A \otimes C) - [(A \times B) + (A \times C)]| \leq 2\varepsilon\beta^{e_1+e_3+1} + \varepsilon\beta^{e_1+e_2} + \varepsilon\beta^{e_1+e_3}$$

$$\leq \varepsilon\beta^{e_1}(2\beta^{e_3+1} + \beta^{e_2} + \beta^{e_3}).$$

Theorem 3.5: Let $A = a\beta^{e_1}$, $B = b\beta^{e_2}$, and $C = c\beta^{e_3}$ with $e_3 \geq e_2$. If $B \oplus C = d\beta^{e_4}$ and if $e_4 \leq e_2$, then the sequence of operations $A \otimes (B \oplus C)$ produces a smaller error bound than the sequence $(A \otimes B) \oplus (A \otimes C)$.

Proof: By Theorem 3.3,

$$|A \otimes (B \oplus C) - A \times (B + C)| \leq \varepsilon\beta^{e_1+e_4} + |A|\, 2\varepsilon\beta^{e_3+1}$$

and by Theorem 3.4,

$$|(A \otimes B) \oplus (A \otimes C) - [(A \times B) + (A \times C)]| \leq \varepsilon\beta^{e_1}(2\beta^{e_3+1} + \beta^{e_2} + \beta^{e_3}).$$

Now

$$\varepsilon\beta^{e_1+e_4} + |A|\ 2\varepsilon\beta^{e_3+1} = \varepsilon\beta^{e_1+e_4} + |a|\ 2\varepsilon\beta^{e_1+e_3+1}$$

$$= \varepsilon\beta^{e_1+e_2}\,(\beta^{e_4-e_2} + |a|\ 2\beta^{e_3-e_2+1})$$

$$\leq \varepsilon\beta^{e_1+e_2}\,(\beta^{e_4-e_2} + |a|\ 2\beta)$$

since $e_3 \geq e_2$. But since $|a| < 1$,

$$\varepsilon\beta^{e_1+e_4} + |A|\ 2\varepsilon\beta^{e_3+1} \leq \varepsilon\beta^{e_1+e_2}\,(\beta^{e_4-e_2} + 2\beta)\,.$$

Since it is assumed that $e_4 \leq e_2$, then

$$\beta^{e_4-e_2} \leq 1 \quad \text{and}$$

$$\varepsilon\beta^{e_1+e_4} + |A|\ 2\varepsilon\beta^{e_3+1} \leq \varepsilon\beta^{e_1+e_2}\,(1 + 2\beta)$$

$$\leq 2\varepsilon\beta^{e_1+e_2}\,(1 + \beta)$$

$$\leq \varepsilon\beta^{e_1}\,(2\beta^{e_2} + 2\beta^{e_2+1})\,.$$

But it was assumed that $e_3 \geq e_2$. Hence

$$\varepsilon\beta^{e_1+e_2} + |A|\ 2\varepsilon\beta^{e_3+1} \leq \varepsilon\beta^{e_1}\,(\beta^{e_2} + \beta^{e_3} + 2\beta^{e_3+1})\,.$$

But $\varepsilon\beta^{e_1}\,(2\beta^{e_3+1} + \beta^{e_2} + \beta^{e_3})$ was the error bound for the sequence $(A \otimes B) \oplus (A \otimes C)$ found in Theorem 3.4. Hence the bound for the operations $A \otimes (B \oplus C)$ is the smaller.

Thus when using pseudo operations to find $A \cdot (B + C)$, if the

$$\text{exponent of } (B \oplus C) \leq \text{Minimum } \{\text{exponent of } B, \text{ exponent of } C\}$$

(which is the case when B and C are nearly equal but of different signs), Theorem 3.5 indicates that the error bound will be smaller if the sequence of operations $A \otimes (B \oplus C)$ is used. This is true regardless of the magnitude of A.

CHAPTER IV

POLYNOMIAL DEFLATION

Iterative methods are commonly used in digital computers to find zeros of a polynomial. Iteration is initiated with a value which is more or less arbitrary and, after a few iterations, a value sufficiently near one of the zeros is reached and the process "homes" in on that zero. When an approximate zero, $\bar{r}$, has been determined in this way, the polynomial is divided by $(x - \bar{r})$ and iteration is continued with the quotient polynomial. Proceeding in this way, all zeros are ultimately determined and the danger of converging twice to the same zero is avoided.

However, the process of dividing the polynomial by $(x - \bar{r})$ can introduce errors in the coefficients of the quotient polynomial. Then the zeros of the quotient polynomial may not be zeros of the original polynomial. To examine the effect of this deflation process, a second degree polynomial will be examined.

Let

$$P(x) = a_2x^2 + a_1x + a_0, \quad a_2 \neq 0,$$

be a polynomial with real coefficients whose zeros are the real numbers r_1 and r_2. Suppose that $\bar{r}_1$, an approximation of the zero r_1, has been found and that $P(x)$ is divided by $(x - \bar{r}_1)$ using synthetic division. The coefficients of the resulting quotient polynomial

$$Q^*(x) = b_1^* x + b_o^*$$

are

$$b_1^* = a_2 , \quad \text{and}$$

$$b_o^* = \bar{r}_1 \circledast a_2 \oplus a_1 .$$

This is an approximation of the polynomial

$$Q(x) = b_1 x + b_o, \quad \text{where}$$

$$b_1 = a_2 \quad \text{and}$$

$$b_o = r_1 a_2 + a_1 ,$$

which has as its zero r_2.

Let $\eta_i = b_i - b_i^*$ $i = 0, 1$. Then

$$\eta_1 = 0 \quad \text{and}$$

$$\eta_o = b_o - b_o^*$$

$$= (r_1 a_2 + a_1) - (\bar{r}_1 \circledast a_2 \oplus a_1) .$$

Let $r_2 + \delta$ be the zero of $Q^*(x)$. Then δ is the error in the second zero of $P(x)$ which results from using $Q^*(x)$ instead of $Q(x)$. Then

$$Q^*(r_2 + \delta) = b_1^*(r_2 + \delta) + b_o^*$$

$$= 0 .$$

But

$$b_1^*(r_2 + \delta) + b_o^* = b_1(r_2 + \delta) + (b_o - \eta_o)$$

$$= b_1 r_2 + b_o + b_1\delta - \eta_o$$

$$= Q(r_2) + b_1\delta - \eta_o \ .$$

But r_2 is a zero of $Q(x)$. Hence

$$b_1^*(r_2 + \delta) + b_o^* = b_1\delta - \eta_o$$

and

$$b_1\delta - \eta_o = 0 \ .$$

Thus the error in the second zero of $P(x)$ is

$$\delta = \frac{\eta_o}{b_1}$$

$$= \frac{\eta_o}{a_2} \ .$$

To establish a bound on the error, it is only necessary to examine η_o.

$$\eta_o = (r_1 a_2 + a_1) - (\bar{r}_1 \otimes a_2 \oplus a_1)$$

and

$$|\eta_o| \leq |(r_1 a_2 + a_1) - (\bar{r}_1 a_2 + a_1)| + |(\bar{r}_1 a_2 + a_1)$$

$$- (\bar{r}_1 \otimes a_2 + a_1)| + |(\bar{r}_1 \otimes a_2 + a_1) - (\bar{r}_1 \otimes a_2 \oplus a_1)|$$

$$\leq |a_2| \ |r_1 - \bar{r}_1| + \varepsilon\beta^{e_{\bar{r}_1} \otimes a_2} + 2\varepsilon\beta^{e_{\bar{r}_1} \otimes a_2 + e_{a_1} + 1}$$

where

$e_{\bar{r}_1 \circledast a_2}$ is the exponent of $\bar{r}_1 \circledast a_2$ and

e_{a_1} is the exponent of a_1.

Then

$$|\delta| \leq \frac{|a_2|\ |\bar{r}_1 - r_1| + \varepsilon\beta^{e_{\bar{r}_1 \circledast a_2}}\left(1 + 2\beta^{e_{a_1}+1}\right)}{|a_2|}. \qquad (4.1)$$

If $\bar{r}_1$ is an M-digit approximation of r_1 which differs with r_1 in the M^{th} digit, then

$$|\bar{r}_1 - r_1| \leq \beta^{-M}\beta^{e_{\bar{r}_1}}$$

and

$$|\delta| \leq \frac{|a_2|\ \beta^{-M}\beta^{e_{\bar{r}_1}} + \varepsilon\beta^{e_{\bar{r}_1 \circledast a_2}}\left(1 + 2\beta^{e_{a_1}+1}\right)}{|a_2|}$$

$$\leq \frac{|a_2|\ 2\varepsilon\beta^{e_{\bar{r}_1}} + \varepsilon\beta^{e_{\bar{r}_1 \circledast a_2}}\left(1 + 2\beta^{e_{a_1}+1}\right)}{|a_2|}.$$

But

$$e_{\bar{r}_1 \circledast a_2} \leq e_{\bar{r}_1} + e_{a_2}$$

and

$$|\delta| \leq \frac{\varepsilon\beta^{e_{\bar{r}_1}}\left(2|a_2| + \beta^{e_{a_2}} + 2\beta^{e_{a_1}+e_{a_2}+1}\right)}{|a_2|}. \qquad (4.2)$$

It is apparent from Equation 4.2 that the error bound for the second zero is directly affected by the magnitude of the first zero. Therefore, in using the deflation process on a second degree polynomial, the best procedure would be to find the zero with the smallest absolute magnitude first. This procedure would produce a smaller error bound for the second zero.

BIBLIOGRAPHY

1. Ashenhurst, R. L. and Metropolis, N., "Unnormalized Floating Point Arithmetic," Journal of the Association for Computing Machinery, Vol. 6, 1959, pp. 415-428.

2. Birkhoff, G. and MacLane, S., A Survey of Modern Algebra, New York: The MacMillian Company, 1941.

3. Carr, John W., III, "Error Analysis in Floating Point Arithmetic," Communications of the Association for Computing Machinery, Vol. 2, 1959, pp. 10-15.

4. Chu, Yaohan, Digital Computer Design Fundamentals, New York: McGraw-Hill Book Company, Inc., 1962.

5. Goldstein, H. H. and Von Neumann, J., "Numerical Inverting of Matrices of High Order," Bulletin of the American Mathematical Society, Vol. 53, 1947, pp. 1021-1099.

6. Householder, Alston S., "Generation of Errors in Digital Computation," Bulletin of the American Mathematical Society, Vol. 60, 1954, pp. 234-247.

7. Householder, Alston S., Principles of Numerical Analysis, New York: McGraw-Hill Book Company, Inc., 1953.

8. Landau, E. G. H., Foundations of Analysis: The Arithmetic of Whole, Rational, Irrational and Complex Numbers. Trans. F. Steinhardt, New York: Chelser Publishing Company, 1951.

9. Olmsted, John M. H., The Real Number System, New York: Appleton-Centry-Crofts, 1962.

10. Scarborough, J. B., Numerical Mathematical Analysis, Baltimore: The John Hopkins Press, 5th Edition, 1962.

11. Wadey, W. G., "Floating-Point Arithmetics," Journal of the Association for Computing Machinery, Vol. 7, 1960, pp. 129-139.

12. Wilkinson, J. H., "Error Analysis of Floating Point Computation," Numerische Mathematik, Vol. 2, 1960, pp. 319-340.

13. Wilkinson, J. H., "The Evaluation of Zeros of Ill-Conditioned Polynomials, Part I," Numerische Mathematik, Vol. 1, 1959, pp. 150-166.

14. Wilkinson, J. H., Rounding Errors in Algebraic Processes, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.